Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-Level Learning Framework

Yiqun Xie[1*], Erhu He[2*], Xiaowei Jia[2], Weiye Chen[1], Sergii Skakun[1], Han Bao[3], Zhe Jiang[4], Rahul Ghosh[5], Praveen Ravirathinam[5]

[1]University of Maryland; [2]University of Pittsburgh; [3]University of Iowa; [4]University of Florida; [5]University of Minnesota
xie@umd.edu, {erh108,xiaowei}@pitt.edu, {weiyec, skakun}@umd.edu, han-bao@uiowa.edu, zhe.jiang@ufl.edu, {ghosh128,pravirat}@umn.edu

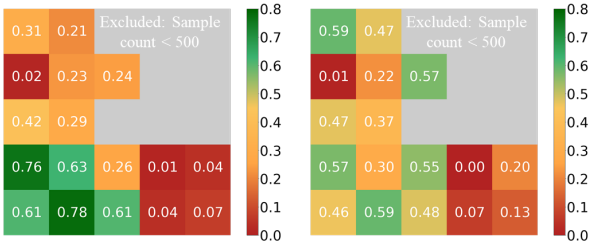
Abstract

Fairness related to locations (i.e., "where") is critical for the use of machine learning in a variety of societal domains involving spatial datasets (e.g., agriculture, disaster response, urban planning). Spatial biases incurred by learning, if left unattended, may cause or exacerbate unfair distribution of resources, social division, spatial disparity, etc. The goal of this work is to develop statistically-robust formulations and model-agnostic learning strategies to understand and promote spatial fairness. The problem is challenging as locations are often from continuous spaces with no well-defined categories (e.g., gender), and statistical conclusions from spatial data are fragile to changes in spatial partitionings and scales. Existing studies in fairness-driven learning have generated valuable insights related to non-spatial factors including race, gender, education level, etc., but research to mitigate location related biases still remain in its infancy, leaving the main challenges unaddressed. To bridge the gap, we first propose a robust space-as-distribution (SPAD) representation of spatial fairness to reduce statistical sensitivity related to partitioning and scales in continuous space. Furthermore, we propose a new SPAD-based stochastic strategy to efficiently optimize over an extensive distribution of fairness criteria, and a bi-level training framework to enforce fairness via adaptive adjustment of priorities among locations. Experiments and case studies on real-world agricultural monitoring show that SPAD can effectively reduce sensitivity in spatial fairness evaluation and the proposed stochastic bi-level training framework can greatly improve the fairness.

Introduction

The goal of spatial fairness, or fairness by "where", is to reduce biases that has significant linkage to the locations or geographical areas of data samples. Such biases, if left unattended, may cause or exacerbate unfair distribution of resources, social division, spatial disparity, and weaknesses in resilience or sustainability (CNBC 2020).

In the following, we illustrate the societal importance of spatial fairness using an example application context in agriculture. Food production is witnessing tremendous supply stresses as a result of rapidly increasing population, climate change, etc. The urgency of the problem has led to major national and international efforts to monitor crops at large scales (e.g., NASA Harvest, G20's GEO-GLAM global agriculture monitoring initiative), and these systems and alike heavily rely on both satellite Earth-observation imagery and learning methods (Kamilaris et al. 2018; Kussul et al. 2017). More importantly, resulting products such as crop maps and acreage estimates (Olofsson et al. 2014) are used to inform critical actions (e.g., distribution of subsidies (NASEM 2018; Bailey and Boryan 2010; Boryan et al. 2011)) to mitigate risks (e.g., natural disturbance incurred food shortage) and support local farmers, which are necessary for continued sustainability and stability. However, current products used to support these important decisions are largely subject to unfairness across locations. For example, prediction accuracy in one region can be easily compromised to pursue better results at other places (e.g., Fig. 1). Such spatial bias can be especially hurtful for a larger number of small holders who represent the main production force behind minor crops (CNBC 2020; USDA 2021; Waldner et al. 2019). Similarly, they can lead to unfair damage estimations (e.g., decrease in yield) due to floods, drought and hurricanes, which are often used to calculate farm insurance. Broadly, spatial fairness has important implications in decision-making across many domains, including disaster management (e.g., floods, wildfires), large-scale carbon monitoring which affects carbon tax, transportation (e.g., traffic and accident prediction, delivery estimation, demand forecast), and many more.



(a) Global F1: 0.573 (b) Global F1: 0.572

Figure 1: Spatial bias examples. (a) and (b) show F1-scores of tomato classification by the same model (trained twice).

The formulation and enforcement of spatial fairness introduce several major challenges. First, unlike traditional categorical-attribute-based fairness (e.g., race or gender-based), spatial domain is a continuous space, which means the "categories" are not well-defined or given-for-free. Sec-

ond, statistics (e.g., fairness scores based on variance) calculated from spatial datasets are fragile or sensitive to both the partition of space and scales, which is also known as the modifiable areal unit problem (MAUP; detailed in Def. 2). In other words, conclusions on "fair" or "unfair" can be easily altered by simple changes in spatial partitions or scales. The lack of consideration on MAUP has led to major societal concerns such as the recent debate on partisan gerrymandering at the US Supreme Court (NPR 2019).

Despite the importance of spatial fairness for the use of deep learning in societal applications, research on this topic is still in its infancy and has barely been studied explicitly in the context of deep learning. Traditional line of research on fairness and equity in space mainly focuses on direct analysis over existing maps or their derivatives (e.g., COVID-19 statistics, access to resources) (Karaye and Horney 2020; Thebault-Spieker, Hecht, and Terveen 2018; Thebault-Spieker, Terveen, and Hecht 2017), which does not aim to address spatial fairness issues entangled with machine learning or deep learning techniques, i.e., improving the techniques' ability to preserve spatial fairness in training or prediction. Extensive learning-based fairness research has been conducted, which is largely focused on pre-defined categorical-attribute-based fairness (e.g., race and gender), including regularization (Zafar et al. 2017; Yan and Howe 2019; Kamishima, Akaho, and Sakuma 2011; Serna et al. 2020), sensitive category de-correlation (Sweeney and Najafian 2020; Zhang and Davidson 2021; Alasadi, Al Hilli, and Singh 2019), data collection/filtering strategies (Jo and Gebru 2020; Yang et al. 2020; Steed and Caliskan 2021), and more (e.g., a recent survey (Mehrabi et al. 2021)). These fairness-aware methods have been used for tasks related to face detection (Serna et al. 2020; Alasadi, Al Hilli, and Singh 2019), language processing (Sweeney and Najafian 2020; Cho et al. 2021), online bidding (Nasr and Tschantz 2020; Ilvento, Jagadeesan, and Chawla 2020), etc. However, existing formulations and methods have yet to address the new challenges brought by spatial fairness, where conclusions can be easily flipped due to statistical sensitivity introduced by MAUP. Finally, heterogeneity-aware learning (Xie et al. 2021a,b) automatically captures differences in data distributions in space, but has not considered fairness.

We aim to tackle the challenges by exploring new formulations and model-agnostic learning frameworks that are spatially-explicit and statistically-robust. Specifically, our contributions are:

- We propose a SPace-As-Distribution (SPAD) representation to formulate and evaluate spatial fairness in the context of continuous space, which mitigates the statistical sensitivity problems introduced by MAUP.
- We propose a SPAD-based stochastic strategy to efficiently optimize over an extensive distribution of candidate criteria for spatial fairness, which are needed to harness MAUP.
- We propose a bi-level player-referee training framework to enhance spatial fairness enforcement via adaptive adjustments of training priorities among locations.

Experiments on real datasets show that the proposed SPAD-based formulation and stochastic training can effectively promote fairness with improved robustness against MAUP-incurred sensitivity. The bi-level training also improves the stability of the model and fairness results compared to traditional regularization-based paradigms.

Key Concepts

Definition 1 ***Partition*** p ***vs. Partitioning*** $\mathcal{P}$**.** *In this paper, a partitioning* $\mathcal{P}$ *splits an input space into* m *individual partitions* p_i*, i.e.,* $\mathcal{P} = \{p_1, ..., p_i, ..., p_m\}$.

Definition 2 ***Modifiable Areal Unit Problem (MAUP).*** *MAUP states that statistical results and conclusions are sensitive to the choice of space partitioning* $\mathcal{P}$ *and scale. A change of scale (e.g., represented by the average area of* $\{p_i \,|\, \forall p_i \in \mathcal{P}\}$*) always infers a change of* $\mathcal{P}$ *but not vice versa. MAUP is often considered as a dilemma as statistical results are expected to vary if different aggregations or groupings of locations are used.*

Definition 3 ***Fairness measure*** M_{fair}**.** *A statistic used to evaluate the fairness across a learning model's performance across several mutually-exclusive groups of individuals. For example,* M_{fair} *can be variance of accuracy across groups. In this paper, groups are defined by partitions* $p \in \mathcal{P}$.

Within the scope of this work, we consider partitionings $\mathcal{P}$ that follow a $s_1 \times s_2$ pattern (i.e., s_1 rows by s_2 columns). Fig. 2 shows an illustrative example of the effect of MAUP on spatial fairness evaluation. Fig. 2 (a1) and (b1) show two example spatial distributions of prediction results (green: correct; red: wrong): (a1) has a large bias where the left side has 100% accuracy and the right side has 0%, and (b1) has a reasonably even distribution of each. However, as shown in Fig. 2 (a2-3) and (b2-3), different partitionings or scales can lead to completely opposite conclusions, making fairness scores fragile in the spatial context.

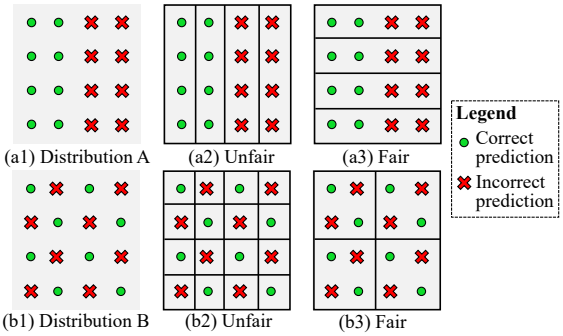


Figure 2: Illustrative examples showing sensitivity to both space-partitioning and scale.

Formulation and Method

In this section, we first propose a novel space-as-distribution (SPAD) formulation to mitigate MAUP-incurred statistical sensitivity for fairness evaluation. Then, we propose a SPAD-based stochastic strategy as well as a bi-level training framework to enforce spatial fairness for an input deep network $\mathcal{F}$ selected by users.

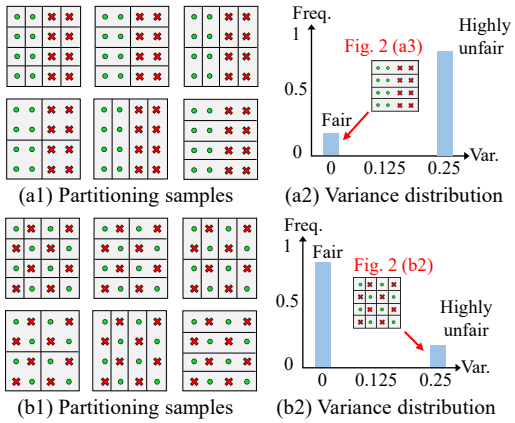


Figure 3: Distributional representation by SPAD.

Space as a Distribution of Partitionings

As grouping of locations is naturally needed for fairness evaluation using common performance metrics (e.g., precision, recall, accuracy), in the scope of this work we focus on scenarios where space-partitionings $\mathcal{P}$ are used to generate location groups; in other words, each partition $p \in \mathcal{P}$ is analogical to a gender, race, etc. in related fairness studies. However, due to the MAUP dilemma (Def. 2), conclusions drawn from most – if not all – of common statistical measures are fragile to the variability in space-partitionings and scales. If this issue is ignored, then one may unintentionally or intentionally introduce additional bias (e.g., partisan gerrymandering (NPR 2019)).

Thus, instead of relying on fragile scores calculated from a fixed partitioning or scale, we propose a SPace-As-Distribution (SPAD) representation to define spatial fairness. The idea is to go beyond a single partitioning or scale by treating space-partitionings at different scales $\{\mathcal{P}\}$ as outcomes of a generative process governed by a statistical distribution. As mentioned in key concepts, in this work we consider partitionings that follow a pattern of s_1 rows by s_2 columns. So, in this case, an example generative process may follow a joint two-dimensional distribution $Prob(s_1, s_2)$ where $s_1,\ s_2 \in \mathbb{Z}^+$, $s_1 \leq row_{max}$, $s_2 \leq col_{max}$ (e.g., 10). By default, one may assume a uniform distribution where $Prob(s_1, s_2) = (row_{max} \cdot col_{max})^{-1}$ (for equal-size partitioning), but this scheme also allows users to flexibly impose a different distribution or prior, which may be dynamically adjusted based on intermediate results.

With the SPAD representation, spatial fairness becomes a distribution of scores, which can more holistically reflect fairness situations across a diverse set of partitions and scales. As an example, Fig. 3 (a1) and (b1) show the same set of partitioning samples (different patterns and scales) overlaid on top of distributions A and B in Fig. 2, respectively. The variance of accuracy across partitions for all 6 partitioning samples are aggregated in (a2) and (b2), where lower variance means fairer results. As we can see, with the distributional extension, the majority of scores reflect our expected results on the fairness evaluation for distributions A and B, and the partitioning samples leading to unexpected results become outliers (highlighted by red arrows).

Once a distribution of scores is obtained from the SPAD representation, summary statistics can be conveniently used for fairness evaluation based on application preferences (e.g., mean). Finally, with SPAD, the formal formulation of spatial-fairness-aware learning is defined as follows:

$$\min_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Gamma}} Prob(\boldsymbol{\Gamma}) \cdot M_{fair}(\mathcal{F}_{\boldsymbol{\Theta}}, M_{\mathcal{F}}, \mathcal{P}_{\boldsymbol{\Gamma}}) d_{\boldsymbol{\Gamma}} \tag{1}$$

where $\mathcal{F}$ is an input deep network with parameters $\boldsymbol{\Theta}$; $\boldsymbol{\Gamma}$ parameterizes a space-partitioning $\mathcal{P}$ (e.g., number of rows and columns for $s_1 \times s_2$-partitionings) that are related to its probability $Prob(\cdot)$ as specified by a statistical distribution (e.g., uniform or user-defined); $M_{\mathcal{F}}$ is a metric used to evaluate the performance of a model $\mathcal{F}$ (e.g., F1-score); and M_{fair} is a fairness measure (loss) that is defined as:

$$M_{fair}(\mathcal{F}_{\boldsymbol{\Theta}}, M_{\mathcal{F}}, \mathcal{P}) = \sum_{p \in \mathcal{P}} \frac{d(M_{\mathcal{F}}(\mathcal{F}_{\boldsymbol{\Theta}}, p), E_{\mathcal{P}})}{|\mathcal{P}|} \tag{2}$$

where p is a partition in $\mathcal{P}$ (Def. 1), $d(\cdot, \cdot)$ is a distance measure (e.g., squared or absolute distance), $M_{\mathcal{F}}(\mathcal{F}_{\boldsymbol{\Theta}}, p)$ is the score (e.g., F1-score) of $\mathcal{F}_{\boldsymbol{\Theta}}$ on p's training data, $|\mathcal{P}|$ is the number of partitions in $\mathcal{P}$, and $E_{\mathcal{P}}$ is another key variable, which represents the mean (expected) performance at each local partition $p \in \mathcal{P}$. If $M_{\mathcal{F}}(\mathcal{F}_{\boldsymbol{\Theta}}, p)$ has a large deviation from the mean (weighted or unweighted), the model $\mathcal{F}_{\boldsymbol{\Theta}}$ is potentially unfair across partitions. Finally, $E_{\mathcal{P}}$ here is calculated from a based model $\mathcal{F}_{\boldsymbol{\Theta}_0}$, where parameters $\boldsymbol{\Theta}_0$ are trained without any consideration of spatial fairness:

$$E_{\mathcal{P}} = \sum_{p \in \mathcal{P}} \frac{M_{\mathcal{F}}(\mathcal{F}_{\boldsymbol{\Theta}_0}, p)}{|\mathcal{P}|} \tag{3}$$

The benefit of using $\mathcal{F}_{\boldsymbol{\Theta}_0}$ to set the mean is that, ideally, we want to maintain the same level of overall model performance (e.g., F1-score without considering spatial fairness) while improving spatial fairness. Thus, this choice automatically takes the overall model performance into consideration as the objective function (Eq. (1)) will increase if $\mathcal{F}_{\boldsymbol{\Theta}}$'s overall performance diverges too far from it (e.g., a model that yields a 0 F1-scores on all partitions – which is fair but poor – will not be considered as a good candidate).

SPAD-based Stochastic Training

A direct way to incorporate the distributional SPAD representation into the training process – either through loss functions or the bi-level method to be discussed in the next section – is to aggregate results from all the partitionings $\{\mathcal{P}\}$ for each iteration or epoch. However, this is computationally expensive and sometimes prohibitive. For example, the number of possible partitionings can be exponential to data size (e.g., the number of sample locations) when general partitioning schemes are considered (e.g., arbitrary, hierarchical, or $s_1 \times s_2$ partitionings with unequal-size cells). Even for equal-size $s_1 \times s_2$ partitionings, there can be easily over hundreds of candidates when large s_1 and s_2 values (e.g., 10, 40, or more) are used for large-scale applications.

Thus, we propose a stochastic training strategy for SPAD to mitigate the cumbersome aggregation. Considering SPAD as a statistical generative process G, in each iteration or

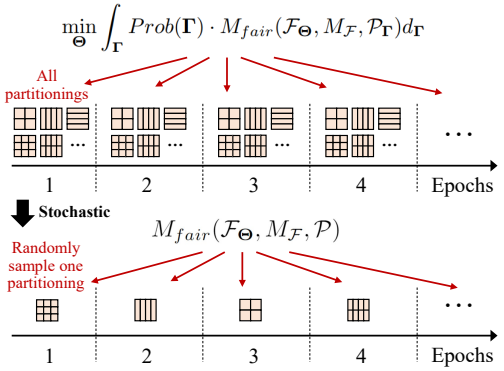


Figure 4: SPAD-based stochastic training strategy.

epoch, we randomly sample a partitioning from G and use it to evaluate fairness-related loss (Def. 3). For example, for equal-size $s_1 \times s_2$ partitionings, each time the generator may randomly sample (s_1, s_2) from a joint discrete distribution (Fig. 4). In this way, the probability of each partitioning (Eq. (1)) is automatically taken into consideration during optimization over epochs. In addition, in scenarios where the difficulty of achieving fairness varies for different partitionings, the SPAD-based stochastic strategy may accelerate the overall convergence. It may first help a subset of partitionings reach good fairness scores faster without the averaging effect, which may in turn help related partitionings to move out local minima traps. In practice, we have three further recommendations for implementation:

- **Unconstrained initial training:** Ideally, we wish to maintain a high overall performance (e.g., F1-scores) while improving fairness across locations. However, it can be premature to try to find a balance between the two objectives when the model still has a very poor overall performance (e.g., untrained). Hence, we keep fairness-related losses or constraints on-hold at the beginning, and optimize parameters by pure prediction errors till stable.
- **Epoch as a minimum unit:** Deep network training often involves mini-batches (i.e., a middle-ground between stochastic and batch gradient descent). As a result, the combined randomness of mini-batches and SPAD-based stochastic strategy may make the training unstable. Thus, using epoch as a minimum unit for changing partitioning samples can help reduce the superposed randomness.
- **Increasing frequency:** Extending the last point, denote k as the number of continuous epochs to train before a partitioning sample is changed. At the beginning of training, a biased model without any fairness consideration may need more epochs to make meaningful improvements, which means a larger k (e.g., 10) is preferred. In contrast, towards the end of the training, a large k can be undesirable as it may cause the model to overfit to a single partitioning at the finish. Thus, we recommend a decreasing k (finally $k = 1$) during training.

Bi-level Fairness Enforcement

A traditional way to incorporate fairness loss (e.g., Eq. (2)) is to add it as a term in the loss function, e.g., $\mathcal{L} = \mathcal{L}_{pred} + \lambda \cdot M_{fair}$, where $\mathcal{L}_{pred}$ is the prediction loss (e.g., cross-entropy or dice loss) and λ is a scaling factor or weight. This regularization-based formulation has three limitations when used for spatial-fairness enforcement: (1) Since deep learning training often uses mini-batches due to data size, it is difficult for each mini-batch to contain representative samples from all partitions $\{p_i \,|\, \forall p_i \in \mathcal{P}\}$ when calculating M_{fair}. (2) To reflect true fairness over partitions, metrics $M_{\mathcal{F}}$ used in M_{fair} in Eq. (2) are ideally exact functions such as precision, recall or F1-scores. However, since many of the functions are not differentiable as a loss function (e.g., with the use of $\arg\max$ to extract predicted classes), approximations are often needed (e.g., threshold-based, soft-version), which introduce extra errors. Additionally, as such approximations are used to further derive fairness indicators (e.g., $M_{\mathcal{F}}$), the uncertainty created by the errors can be quickly accumulated and amplified; and (3) The regularization term M_{fair} requires another scaling factor λ, the choice of which directly impacts final output and varies from problem to problem.

To mitigate these concerns, we propose a bi-level training strategy that disentangles the two types of losses with different purposes (i.e., $\mathcal{L}_{pred}$ and M_{fair}). Specifically, there are two levels of decision-making in-and-between epochs:

- **Partitioning-level ($\mathcal{P}$):** Before each epoch, a referee evaluates the spatial fairness using Eq. (2) with exact metrics $M_{\mathcal{F}}$ (e.g., F1-score); no approximation is needed as back-propagation is not part of the referee. The evaluation is performed on all partitions $p_i \in \mathcal{P}$, guaranteeing the representativeness. Note that the model is evaluatable for the very first epoch because the fairness-driven training starts from a base model, as discussed in the previous section and explanations for Eq. (2). Based on an individual partition p_i's deviation $d(M_{\mathcal{F}}(\mathcal{F}_{\Theta}, p_i), E_{\mathcal{P}})$ (a summand in M_{fair}'s numerator in Eq. (2)), we assign its learning rate η_i for this epoch as:

$$\eta_i = \frac{\eta'_i - \eta'_{min}}{\eta'_{max} - \eta'_{min}} \cdot \eta_{init} \tag{4}$$

$$\eta'_i = \max(-(M_{\mathcal{F}}(\mathcal{F}_{\Theta}, p_i) - E_{\mathcal{P}}), 0) \tag{5}$$

  where η_{init} is the learning rate used to train the base model, $\eta'_{min} = \arg\min_{\eta'_i}\{\eta'_i \,|\, \eta'_i > 0, \forall i\}$, and $\eta'_{max} = \arg\max_{\eta'_i}\{\eta'_i \,|\, \forall i\}$.

  The intuition is that, if a partition's fairness measure is lower than the expectation E_p, its learning rate η_i will be increased (relatively to other partitions') so that its prediction loss will have higher impact during parameter updates in this epoch. In contrast, if a partition's performance is the same or higher than the expectation, its η_i will be set to 0 to prioritize other lower-performing partitions. Positive learning rates after the update are normalized back to the range $[0, \eta_{init}]$ to keep the gradients more stable. This bi-level design also relieves the need for an extra scaling factor to combine the prediction and fairness losses.
- **Partition-level (p):** Using learning rates $\{\eta_i\}$ assigned by the referee, we perform regular training with the prediction loss $\mathcal{L}_{pred}$, iterating over data in all individual partitions $p_i \in \mathcal{P}$ in mini-batches.

Dataset and Implementation Details

Dataset: Accurate mapping of crops is critical for estimating crop areas and yield, which are often used for distributing subsidies and providing farm insurance over space. Our input **X** for crop and land cover classification is the multi-spectral remote sensing data from Sentinel-2 in Central Valley, California, and the study region has a size of 4096×4096 ($\sim$6711 km^2 at 20m resolution). We use the multi-spectral data captured in August, 2018 for the mapping, and each location has reflectance values from 10 spectral bands, which are used as input features. The label **y** is from the USDA Crop Data Layer (CDL) (CDL 2017). In our tests, we randomly select 20%, 20%, and 60% locations for training, validation and testing, respectively.

Implementation details: As mentioned in scope, we consider $s_1 \times s_2$ partitionings. In experiments, to allow comparisons with non-stochastic-based SPAD methods (computationally expensive), we set the maximum values for s_1 and s_2 to 5, which leads to 24 different equal-size partitionings (the 1×1 partitioning is excluded).

We use an 8-layer deep neural network (DNN) as a base model to test the proposed SPAD method; SPAD does not assume specific network architectures. The DNN model takes inputs of multi-spectral data at each location and outputs the land cover label. In our experiment, we first train an initial DNN model for 300 epochs (converged) without considering the fairness, using Adam ($\alpha = 0.001$) as the optimizer. From this base model, we further implement different candidate approaches to improve fairness (variants with no base model are also considered). Based on the strategy discussed in stochastic training, at the beginning of fairness training, we keep each sampled partitioning for 10 epochs before moving onto the next, and iterate over 48 different samples (i.e., can be interpreted as two full enumerations over all 24 partitioning candidates). In the middle stage, we keep each partitioning for 5 epochs, and iterate over 96 samples (i.e., similar to four full enumerations). Finally, each epoch will sample a new partitioning, which continues for 240 samples. Overall there are 50 expected epochs for each partitioning.

Both weighted and unweighted F-1 scores are considered as the performance metric $M_{\mathcal{F}}$ in Eqs. (2) and (3).

Experiments

Our experiments aim to answer the following questions: [1]

- **Q1**. Does the SPAD representation improve spatial fairness over different space-partitionings?
- **Q2**. Does the bi-level training strategy improve over regularization-based approaches?
- **Q3**. Is the SPAD-based stochastic training able to maintain or improve fairness with smaller computational load?
- **Q4**. Can the proposed approach help reach a fairer solution while maintaining a similar level of overall/global performance? Does training from an unconstrained base model (no fairness consideration) help reach this goal?

[1] Additional results and code are included in the supplementary file.

The results to these questions can serve as an initial base for spatial-fairness driven learning. Based on the questions, our candidate methods are:

- **Base:** The base deep learning model (8-layer DNN) without consideration of spatial fairness.
- **Single:** Spatial fairness is evaluated and improved using a single space-partitioning $\mathcal{P}$. Specifically, our experiment includes Single-(1,4) and Single-(4,1), which use 1×4 and 4×1 partitionings, respectively.
- **REG:** Spatial fairness is enforced using the SPAD representation by adding a regularization term to the loss function. As F1-score is not differentiable, we use standard approximation via the threshold-based approach, which amplifies softmax predictions $\hat{y}$ over a threshold γ to 1 to suppresses others to 0 using $1 - \text{ReLU}(1 - A \cdot \text{ReLU}(\hat{y} - \gamma))$, where A is a sufficiently large number ($A = 10000$ in our tests; more details in the supplementary file). The scaling factor λ for the regularizer is set to 5.
- **SPAD:** The proposed approach using the SPAD representation with the stochastic and bi-level training strategies.
- **SPAD-GD:** SPAD without the stochastic strategy, which aggregates over gradients from all 24 partitionings before making parameter updates in each round.
- **SPAD-no-base:** SPAD that starts training without using an unconstrained base model (explained in the stochastic training section). Since here we do not have a ready-to-use expected performance ($E_{\mathcal{P}}$ in Eq. (3)) from the base model, we randomly initiate $E_{\mathcal{P}}$ and dynamically update it with the new learned parameters in each epoch.
- **SPAD-10-eps**: In the stochastic training, this version keeps using each sampled partitioning for $k = 10$ epochs, without decreasing k to 1 near the end, which may make the model biased towards the last sample (explained in the stochastic training section).

Comparison to the regularization-based method

We compare the fairness achieved by SPAD, the base DNN model (without considering fairness) and the REG method in Fig. 5. For each partitioning $\mathcal{P}$ (x-axis), we report the mean of the absolute distances between F1-scores achieved on each partition p and the average performance over all partitions $\{p \in \mathcal{P}\}$; both weighted and unweighted F-1 scores are considered. In Tables 1 and 2, we summarize the overall performance (global F1-scores), the sum of mean absolute distance $S(d)_{\text{mean}}$ and the sum of maximum absolute distance $S(d)_{\text{max}}$ across all partitionings using weighted and unweighted F-1, respectively.

Fig. 5 shows that both SPAD and REG achieve lower mean absolute distances over all space partitionings compared to the base model, confirming the effectiveness of the SPAD representation in improving the fairness (**Q1**). Comparing SPAD and REG, we can see that SPAD consistently outperforms REG in the experiments (**Q2**), which shows that the bi-level design is more effective in enforcing spatial fairness than regularization terms by improving sample representativeness, allowing the use of exact metrics (i.e., no need

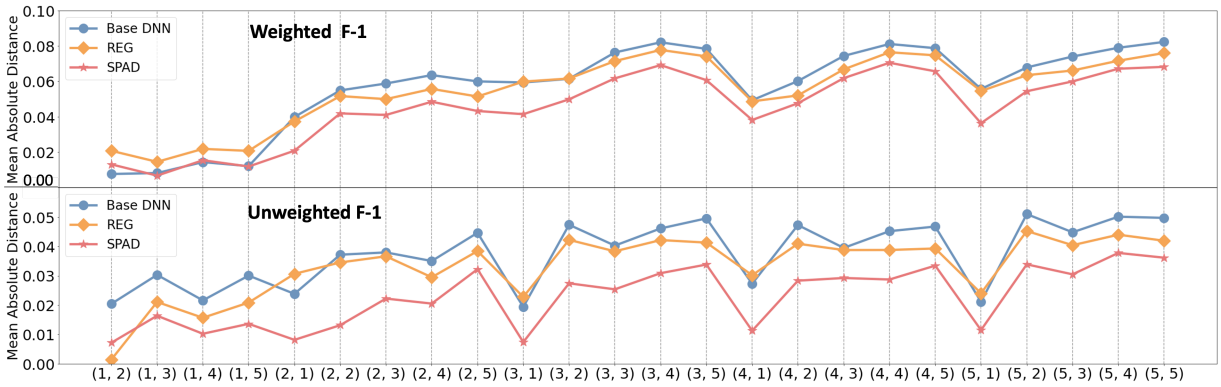


Figure 5: Fairness comparison amongst SPAD, REG and the base model over all the partitionings.

Table 1: Global classification performance and sums of mean/max absolute distances over all partitionings using weighted F1 score as the performance metric (results with a substantially reduced global F1-score is denoted with *).

Method	**Weighted F1**	$S(d)_{\text{mean}}$	$S(d)_{\text{max}}$
Base DNN	0.572	1.379	3.799
REG	0.566	1.319	3.821
Single-(1,4)	0.576	1.356	3.666
Single-(4,1)	0.542	1.355	3.712
SPAD-GD	0.573	1.275	3.571
SPAD-10-eps	0.573	1.186	3.421
SPAD-no-base	0.507*	1.589*	4.595*
SPAD	0.573	**1.094**	**3.185**

Table 2: Global classification performance and sums of mean/max absolute distances over all partitionings using unweighted F1 score as the performance metric (results with a substantially reduced global F1-score is denoted with *).

Method	**Unweighted F1**	$S(d)_{\text{mean}}$	$S(d)_{\text{max}}$
Base DNN	0.377	0.906	1.808
REG	0.381	0.799	1.808
Single-(1,4)	0.362	0.627	1.392
Single-(4,1)	0.368	0.685	1.517
SPAD-GD	0.372	0.602	1.384
SPAD-10-eps	0.361	0.582	1.393
SPAD-no-base	0.318*	0.469*	0.981*
SPAD	0.374	**0.549**	**1.337**

to use approximations of F1-scores for differentiability purposes), and eliminating the need for an extra scaling factor for the regularizer which may add extra sensitivity.

From the first columns of Tables 1 and 2, we can see that SPAD is able to maintain a similar overall/global classification performance compared to the base DNN, which does not have any fairness consideration. Meanwhile, the second and third columns in the tables show that our method can significantly reduce the sums of mean and max absolute distance over all partitionings. This confirms that SPAD can effectively promote the fairness without compromising the classification performance (**Q4**).

Comparison to partitioning-specific SPAD

Next, we aim to verify that SPAD can achieve better fairness over majority of the partitionings compared to non-SPAD-based variants that only only optimizes fairness over a specific spatial partitioning. Fig. 6 shows the fairness performance of partition-specific methods Single-(1,4) and Single-(4,1). The overall trend is that SPAD achieves better spatial fairness in most partitionings by modeling space-partitionings as a distribution (**Q1**). In addition, we can also observe that Single-(4,1) obtains a better fairness result for the given partitioning (4,1), and similarly Single-(1,4) performs better for (1,4). However, their fairness improvements are limited for other partitionings. This conforms to the expectation that partitioning-specific methods are able to reach further improvements on a given $\mathcal{P}$, but cannot generalize well to the others. Tables 1 and 2 (rows 3-4) show the weighted and unweighted F1-scores achieved by Single-(1,4) and Single-(4,1). The numbers confirm that the methods also have similar global F1-scores since our design takes the overall performance into account (Eqs. (2) and (3)). However, they produce larger values of $S(d)_{\text{mean}}$ and $S(d)_{\text{max}}$ (max-absolute-distance figures are in supplementary file), which again confirms the benefits for SPAD.

Interestingly, in both the experiments with weighted and unweigted F1-scores (Fig. 6), SPAD can often get very close to the fairness scores achieved by partitioning-specific methods on their sole-input $\mathcal{P}$ (except for (4,1) in the unweighted case). This shows there are potential dependency relationships between partitionings. We also explored a variant that uses only finer or finest-scale partitionings. One issue we observed is that the method faces difficulty in convergence, leading to both poorer results on fine and coarse scales. This is potentially due to the fact that fairness enforcement at finer-scale naturally leads to stricter criteria. We will examine more effective methods to leverage such potential dependency among partitionings in future work.

Validation of stochastic training strategies

Finally, we validate the effectiveness of the SPAD-based stochastic training strategy (**Q3**). We first compare to the SPAD-GD approach, which aggregates gradients from all partitionings in each epoch. Compared to our SPAD-based stochastic approach, the aggregation in SPAD-GD leads to a heavier computational load and requires longer time for

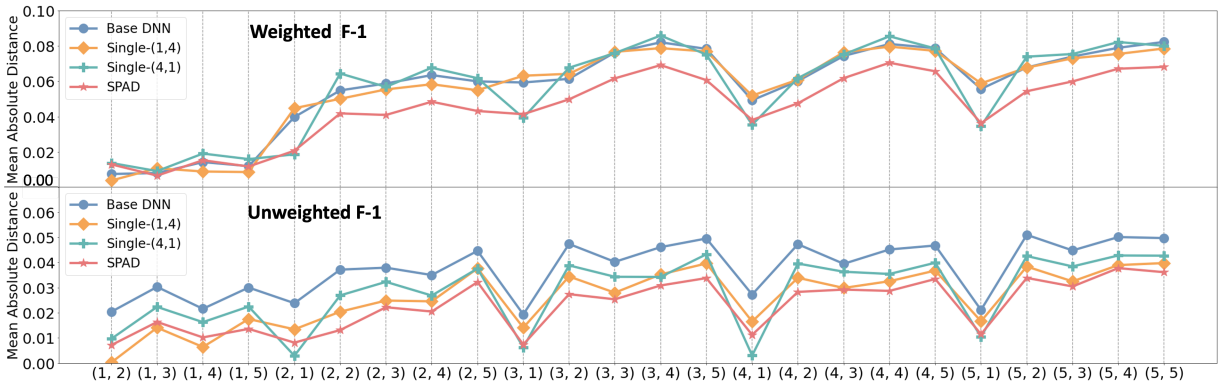


Figure 6: Fairness comparison amongst SPAD, Single-(1,4), Single-(4,1), and the base model over all the partitionings.

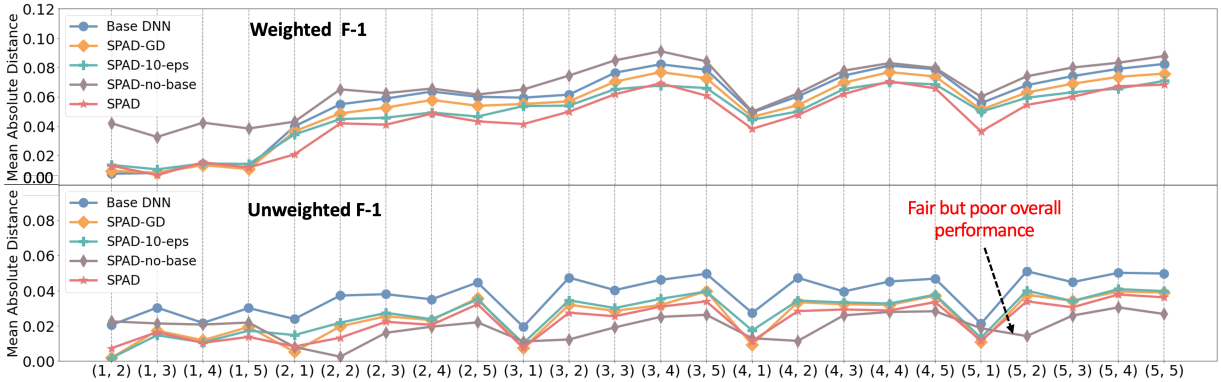


Figure 7: Fairness comparison amongst different optimization methods.

training (i.e., 2.5 hours vs. 9.5 hours using NVIDIA Tesla K80 GPU over two runs). Here we maintain the same number of parameter updates for the two methods, and the only difference is that each SPAD update is made by gradients from a sampled partitioning whereas each SPAD-GD update uses average gradients from all partitionings. Fig. 7 shows their performance comparison. We can see that the two methods have about the same performance for the unweighted scenario (lower part of Fig. 7), which is expected. Interestingly, SPAD outperforms SPAD-GD in the weighted scenario (upper part of Fig. 7). One reason is that the added randomness from the stochastic sampling in SPAD may allow a better chance for the training to move out of local minima traps without the averaging effects, especially when fairness is harder to achieve at the beginning for some partitionings. We also compare to SPAD-10-eps which uses 10 epochs for each sampled partitioning till the end of training. According to Fig. 7, SPAD-10-eps has decreased overall fairness results compared to SPAD. The reason is that SPAD-10-eps, without reducing the epoch number per partitioning, tends to overfit to the last sample partitioning, leading to poorer performance on the rest.

As a stable initial model state is helpful for fairness training, SPAD and other candidate methods start training from a base model (discussed in implementation details). Here we compare SPAD with SPAD-no-base, which enforces fairness right at the start of training. According to its results in Tables 1 and 2 and fairness results in Fig. 7, the method has a substantially reduced global F1-score compared to all other methods (e.g., by 14%), making its fairness results not as interesting (i.e., fair but poor). This shows that the base model is beneficial in improving fairness while maintaining good global performance. In addition, since SPAD-no-base starts focusing on fairness when weights are still pre-mature, its performance tends to be unstable for fairness as well (e.g., may be lower-ranked in terms of both the global F1 score and fairness scores as shown in Table 1).

Conclusions and Future Work

Understanding and controlling location-related bias are critical for fair resource distribution in many important societal domains including agriculture, disaster management, etc. We proposed a new formulation of spatial-fairness driven learning using the SPAD representation, which harnesses statistical sensitivities in fairness evaluations caused by MAUP. We also proposed SPAD-based stochastic and bi-level training strategies to enforce spatial fairness in learning. Experiment results on real-world agriculture monitoring data confirmed that the proposed approach is effective in improving spatial fairness while maintaining a similar level of overall performance. Code, additional details and results are included in the supplementary document.

In future work, we will explore new sampling strategies to improve the computational efficiency of the approach, and the use of the approach for other related scenarios such as regression and dynamic spatio-temporal tasks. We will also expand the experiments using data from more domains (e.g., carbon monitoring) and more base architectures.

Acknowledgments

Yiqun Xie and Weiye Chen are supported in part by NSF awards 2105133 and 2126474, Google's AI for Social Good Impact Scholars program, and the DRI award at the University of Maryland; Erhu He and Xiaowei Jia are supported in part by USGS award G21AC10207, Pitt Momentum Funds award, and CRC at the University of Pittsburgh; Sergii Skakun is supported in part by NASA LCLUC Award 80NSSC21K0314; Han Bao is supported in part by the ISSSF grant from the University of Iowa, and SAFER-SIM funded by US-DOT award 69A3551747131; and Zhe Jiang is supported in part by NSF awards IIS-1850546, IIS-2008973, CNS-1951974 and OAC-2152085.

References

Alasadi, J.; Al Hilli, A.; and Singh, V. K. 2019. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 19–25.

Bailey, J. T.; and Boryan, C. G. 2010. Remote sensing applications in agriculture at the USDA National Agricultural Statistics Service. Technical report, Research and Development Division, USDA, NASS, Fairfax, VA.

Boryan, C.; Yang, Z.; Mueller, R.; and Craig, M. 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5): 341–358.

CDL. 2017. Cropland Data Layer - USDA NASS. https://geography.wr.usgs.gov/science/croplands/pubs2017.html.

Cho, W. I.; Kim, J.; Yang, J.; and Kim, N. S. 2021. Towards Cross-Lingual Generalization of Translation Gender Bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 449–457.

CNBC. 2020. As small U.S. farms face crisis, Trump's trade aid flowed to corporations. Https://www.cnbc.com/2020/09/02/as-small-us-farms-face-crisis-trumps-trade-aid-flowed-to-corporations.html.

Ilvento, C.; Jagadeesan, M.; and Chawla, S. 2020. Multi-category fairness in sponsored search auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 348–358.

Jo, E. S.; and Gebru, T. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316.

Kamilaris, A.; et al. 2018. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147: 70–90.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650. IEEE.

Karaye, I. M.; and Horney, J. A. 2020. The impact of social vulnerability on COVID-19 in the US: an analysis of spatially varying relationships. *American journal of preventive medicine*, 59(3): 317–325.

Kussul, N.; Lavreniuk, M.; Skakun, S.; and Shelestov, A. 2017. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5): 778–782.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

NASEM. 2018. *Improving crop estimates by integrating multiple data sources*. National Academies Press.

Nasr, M.; and Tschantz, M. C. 2020. Bidding strategies with gender nondiscrimination constraints for online ad auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 337–347.

NPR. 2019. Supreme Court Rules Partisan Gerrymandering Is Beyond The Reach Of Federal Courts. Https://www.npr.org/2019/06/27/731847977/supreme-court-rules-partisan-gerrymandering-is-beyond-the-reach-of-federal-court.

Olofsson, P.; Foody, G. M.; Herold, M.; Stehman, S. V.; Woodcock, C. E.; and Wulder, M. A. 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148: 42–57.

Serna, I.; Morales, A.; Fierrez, J.; Cebrian, M.; Obradovich, N.; and Rahwan, I. 2020. Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*.

Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–713.

Sweeney, C.; and Najafian, M. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 359–368.

Thebault-Spieker, J.; Hecht, B.; and Terveen, L. 2018. Geographic Biases are'Born, not Made' Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, 71–82.

Thebault-Spieker, J.; Terveen, L.; and Hecht, B. 2017. Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction*, 24(3): 1–40.

USDA. 2021. Economic Research Service Farm Resources Regions. https://www.ers.usda.gov/webdocs/publications/42298/32489_aib-760_002.pdf.

Waldner, F.; Chen, Y.; Lawes, R.; and Hochman, Z. 2019. Needle in a haystack: Mapping rare and infrequent crops using satellite imagery and data balancing methods. *Remote Sensing of Environment*, 233: 111375.

Xie, Y.; He, E.; Jia, X.; Bao, H.; Zhou, X.; Ghosh, R.; and Ravirathinam, P. 2021a. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*, 767–776. IEEE.

Xie, Y.; Jia, X.; Bao, H.; Zhou, X.; Yu, J.; Ghosh, R.; and Ravirathinam, P. 2021b. Spatial-Net: A Self-Adaptive and Model-Agnostic Deep Learning Framework for Spatially Heterogeneous Datasets. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 313–323.

Yan, A.; and Howe, B. 2019. Fairst: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 552–555.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zhang, H.; and Davidson, I. 2021. Towards Fair Deep Anomaly Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 138–148.